Praktische Aufgabe - Feature Engineering

In dieser Übung bauen wir auf den bereinigten Daten des Data-Mining-Cup auf und werden uns näher mit dem Erzeugen, Verarbeiten und Bewerten von Features beschäftigen. Um sicher zu gehen, dass die hierzu benötigten Daten zu Verfügung stehen und keine Fehler durch mangelhafte Bereinigung des Datensatzes auftreten, ist eine geeignete Version der Daten online zu finden (\\fs23\lehrveranstaltungen\Big Data Analytics\dmc-2016-cleaned.csv).

1. neue Features generieren

Häufig lassen sich aus vorliegenden Features weiter generieren, welche sich besser für den Data-Mining Prozess eignen. Erzeuge exemplarisch folgende neue kategorische Features:

- den Monat des jeweiligen Jahres der Bestellung ($monthOfYear \in [1, 12]$)
- ob die *quantity* dem Modalwert entspricht ($modeQuantity = 1$ falls ja, sonst 0)
- ob ein Gutschein benutzt wurde ($hasVoucher = 1$ falls ja, sonst 0)
- ob der selbe Artikel (*articleID*) in einer Bestellung (*orderID*) mehr als ein mal vorkommt ($sameArticleInOrder = 1$ falls ja, sonst 0)
- mindestens 2 weitere selbst erdachte kategorische Features

Und folgende numerische Features:

- summierter Preis einer Bestellung ($totalOrderPrice$ = Summe über Preise mit gleicher *orderID*)
- Ersparnis eines Bestellpostens ($saving = rrp \times quantity - price$)
- Spanne zwischen dem günstigsten und dem teuersten Artikel innerhalb einer Bestellung (Preis je Artikel wird hierzu durch $\frac{price}{quantity}$ berechnet, Bestellungen werden über die *orderID* identifiziert)
- Rückgabe-Wahrscheinlichkeit für jeden Kunden ($customerReturnProbability = \frac{sum(returnQuantity)}{sum(quantity)}$ je Kunde)
- mindestens 2 weiter selbst erdachte numerische Features

Tipp: DataFrames verfügen über eine groupby(...) Methode

2. Features bewerten

Da zu viele Features häufig dazu führen, dass Data-Mining Algorithmen hohe Laufzeiten haben oder sogar schlechtere Ergebnisse liefern, ist es oft nötig Feature Selection zu betreiben. Einige Verfahren, so genannte Filtermethoden, nutzen hierzu statistische Maße, um die Features zu bewerten. Solche Verfahren habt ihr bereits in der Vorlesung kennen gelernt. Arbeitet zu den in Aufgabe 1 erstellten Features folgende Schritte durch:

- erstellt sinnvolle Plots jedes Features gegenüber der *returnQuantity* (Barplot, Scatterplot, ...)
- berechnet die X^2 Statistik der kategorischen Features gegenüber der *returnQuantity*
- berechnet für jedes numerische Feature die Pearson-Korrelation gegenüber der *returnQuantity*

Dokumentiert hierzu die Plots und die Statistiken der einzelnen Features im Wiki. Für die selbst erstellten Features ist ebenfalls eine kurze inhaltliche Beschreibung vorteilhaft. Diskutiert mögliche Gründe, warum einige Features besser für die Vorhersage der *returnQuantity* geeignet sind. Worin liegen mögliche Gründe, dass einige Features schwächer sind? Welche Nachteile könnten einige Features bei der Vorhersage für zukünftige Daten haben? Gibt es Auffälligkeiten in den berechneten Statistiken?

3. Dimensions-Reduktion mit PCA

Mittels PCA können vor allem bei hochdimensionalen Daten weniger wichtige Dimensionen gefunden und entfernt werden. Dazu werden diese in einen neuen Vektorraum transformiert. Um dies einmal beispielhaft auszuführen, sind die Dateien pca_1.csv und pca_2.csv online verfügbar. Führt mit beiden folgendes aus:

- plotet die Daten (Spalte x gegen Spalte y)
- führt das für PCA nötige Preprocessing auf x und y aus
- transformiert die Daten mittels PCA (ohne Dimensionen zu entfernen) und plotet das Ergebnis

Die Daten aus pca_2.csv verfügen über eine zusätzliche *class* Spalte. Achtet darauf, dass diese nicht durch das Preprocessing oder PCA verändert wird. Plotet für diesen Datensatz die 2 unterschiedlichen Klassen (*class* ist 0 oder 1) in verschiedenen Farben. Stellt die Plots in das Wiki und diskutiert, welche Unterschiede euch auffallen. Was würde passieren, wenn man die Dimensionsreduktion ausführen würde? Wie kann dieses Problem gelöst werde? Welche weiteren Nachteile kann PCA mit sich bringen (etwa hinsichtlich der Interpretation der Daten)?